Cross-Lingual Retrieval for Hindi

JINXI XU AND RALPH WEISCHEDEL
BBN Technologies

In this paper we describe the evaluation results of applying a cross-lingual retrieval model to retrieve Hindi documents relevant to an English query. Though the technique has been previously applied and evaluated for retrieving Chinese and Arabic documents given an English query, what is new about these experiments is porting the model to Hindi in two weeks' time.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms: Algorithms, Experimentation, Measurement

Additional Key Words and Phrases: Hindi, cross-lingual retrieval

1. INTRODUCTION

Cross-lingual information retrieval (CLIR) is the task of retrieving documents in a second language different than the query language. Research and results in CLIR have blossomed, particularly in the last three years as corpora, query sets, and relevance judgments have become available, e.g., through TREC (http://trec.nist.gov); CLEF (http://clef.iei.pi.cnr.it/); and NTCIR (http://research.nii.ac.jp/~ntcadm/index-en.html). In the first half of 2003, a group of researchers collaborated in a unique porting experiment: trying to obtain data and bring up algorithms for a new language in 29 calendar days. The chosen languages were Cebuano and Hindi. This paper reports on our experience with English as the query language and Hindi as the second language. In fact, we were able to retrieve Hindi documents from English queries within two weeks of announcement of the language for the experiments.

One of the advantages of our approach is that many cross-lingual resources can be combined using a simple probabilistic model of relevance and of term translation. The probability of each possible path from the document through translation to a query term is summed to estimate the probability that the document is relevant to the query. Section 2 reviews the statistical language model used to rank documents. Section 3 states how we dealt with stop words and stemming. Section 4 describes how we estimated term translation probabilities from both manual bilingual lexicons and also from parallel texts in English and Hindi, including "pseudo-parallel" texts generated by a machine translation of the Hindi collection. Section 5 summarizes empirical results. Section 6 notes related work, and Section 7 presents our conclusions.

This research is sponsored by the Defense Advanced Research Projects Agency and managed by SPAWAR under contract N66001-00-C-8008. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the Defense Advanced Research Projects Agency, SPAWAR, or the United States Government.
Authors' address: BBN Technologies, 10 Moulton St., Cambridge, MA 02138.

2. MODEL EMPLOYED

The information retrieval model used for Hindi has been reported previously [Xu, et al. 2001], and has been evaluated formally and demonstrated state-of-the-art performance on Arabic and Chinese sources in the Text Retrieval Conferences (TREC). One of the advantages of our approach is that many cross-lingual resources can be combined using a simple probabilistic relevance model and term translation. At the heart of the algorithm is the following ranking measure:

$$P(Q \mid D) = \prod_{e \in Q} (aP(e \mid GE) + (1-a) \sum_{h \in D} P(h \mid D) P(e \mid h))$$

where e is a query word, Q is the query, D is a document, h a term in the Hindi document, and GE a large collection of general English. The crucial aspect of this ranking function is that, for every query word, the sum of the probability of every path from a document word (in the foreign language) through translation to a query word is estimated. How to estimate translation probabilities is discussed in Section 4.

To minimize the need for training data, we estimated the parameters as follows:

- The parameter a is a constant. We fixed it at 0.3 in this work, based on prior experience.
- In the general English (GE) state, we estimated the probability distribution as follows:

  $$P(e \mid GE) = freq(e, GE) / |GE|$$

  where $freq(e, GE)$ is the frequency of English word e in an English corpus and $|GE|$ is the size of the English corpus. Any large English corpus can be used for this purpose; we used TREC volumes 1-5 of English data.
- In the document state (D), we estimated the probability distribution as follows:

  $$P(h \mid D) = freq(h, D) / |D|$$

  where $freq(h, D)$ is the frequency of a Hindi word h in D and $|D|$ is the length of the document.
- The probability of translation to an English word e given a Hindi word h, $P(e|h)$, depends on h and e only.

3. TOKENIZATION ISSUES FOR HINDI

We considered the following three issues: spelling normalization, removal of stop-words, and stemming. For spelling normalization, we used the standard normalization script provided by the Linguistic Data Consortium (LDC). For stop-word removal, we used the 284 stop-words provided by the University of Massachusetts. For stemming, we used the stemming routine of the University of Maryland, with modifications to handle UTF8 encoding. All three resources were obtained from LDC.

4. TRANSLATION PROBABILITIES

The translation probabilities between English and Hindi words are central to our retrieval model. To retrieve Hindi documents using English queries, the retrieval model needs to know $P(e|h)$, the probability that a Hindi word h is translated to an English word e. Translation probabilities were estimated from a number of sources and were linearly combined. The sources include the following:

A: The master English-Hindi lexicon from LDC, augmented with translations of 425 geopolitical entities (obtained from USC/ISI) and translations of 1,700 named entities (obtained from CMU). After stop-word removal and stemming, 78,000 word pairs were obtained. The translation probabilities were uniformly estimated. That is, if a Hindi word h has n possible English translations e_1 to e_n, then

$$P(e_i|h)=1/n.$$

B: A statistical lexicon from IBM derived from a word-aligned parallel corpus. The lexicon is a list of word pairs (e, h), with joint frequency $f(e, h)$. Translation probabilities were calculated according to

$$P(e \mid h) = \frac{f(e,h)}{f(h)}$$
$$f(h) = \sum_{English\ \ words\ \ x} f(x,h)$$

This source has 172,000 word pairs after stemming and removal of stop-words.

C: A sentence-aligned parallel corpus from UC Berkeley, based on the English and Hindi keyword fields of 37,000 BBC Hindi documents provided by MITRE. This parallel corpus contains around 360,000 words in both languages. Translation probabilities were estimated using GIZA++[1]. Around 116,000 translation pairs were extracted.

D: A sentence-aligned parallel corpus collected by LDC, with 1.5 million English words. In addition, a set of 10,000 English-Hindi sentence pairs collected by John Hopkins University was added to the LDC parallel corpus. Translation probabilities were estimated using GIZA++. Around 1.6 million words pairs were extracted.

E: BBC news articles translated by USC/ISI's machine translation system. It consists of around 2,900 articles with 0.4 million words in English. The translated articles were treated as a pseudo-parallel corpus, from which translation probabilities were extracted using GIZA++. Around 380,000 word pairs were extracted.

All the lexical sources were either posted on LDC's processed resources website, or announced by individual groups (e.g., ISI and Berkeley) and posted on their own web sites. Translation probabilities were linearly combined, using

$$P(e \mid h) = \frac{\sum_{s\in\{A,B,C,D,E)} \lambda(s) P(e \mid h, s)}{\sum_{s\in\{A,B,C,D,E\}} \lambda(s)}$$

where $\lambda(s)$ is the mixture weight assigned to source s. The probabilities were normalized in order to handle missing terms in the sources. The weights were chosen based on a set of 93 development queries on the BBC corpus. The development queries were contri-

[1] Downloaded from http://wwwi6.informatik.rwthaachen.de/web/Software/GIZA++.html

Table I. Hindi CLIR Results for BBN's Submitted Runs

	BBNSLA	BBNSLB	BBNSLC	BBSLD	BBNSLMON
5-docs	0.7867	0.7600	0.8000	0.7867	0.7333
10-docs	0.7267	0.7133	0.7133	0.7267	0.6667
15-docs	0.6933	0.6844	0.6444	0.6711	0.5956
20-docs	0.6400	0.6600	0.6333	0.6400	0.5567

buted by University of Maryland (29 queries) and BBN (64 queries), with relevance judgments from University of Maryland, BBN, and UMass. Specifically, the weights are 1, 2, 1, 3, and 2 for sources *A*, *B*, *C*, *D*, and *E* respectively.

5. RESULTS

Two types of query expansion were used: adding additional English terms and adding Hindi terms to the initial queries. English query expansion was performed on a corpus consisting of around 0.5 million documents from two sources, Economic Times (provided by John Hopkins University) and FBIS. The Economic Times is a newswire source from India. Hindi query expansion was performed on the test corpus itself. In both cases, 30 expansion terms were selected based on their TF.IDF weights from the top 10 documents. The TF.IDF formula is based on Allan et al. [1998]. The expansion terms and the original query terms were weighted as follows:

$$wt(t) = wt_{old}(t) + 0.4 \sum_{\text{Top documents } D} tfidf(t, D)$$

where *D*'s are the top retrieved documents. The weight can be interpreted as the frequency term *t* is expected to appear in the query. Therefore, it is used as an exponent in the retrieval formula.

We submitted four cross-lingual runs and one monolingual run:

- BBNSLA: no query expansion
- BBNSLB: English query expansion
- BBNSLC: Hindi query expansion
- BBNSLD: English and Hindi query expansion
- BBNSLMON: mono-lingual run, no query expansion

All English fields of the topics were used in query formulation. The fields include title, description, and narrative, similar to TREC topics, as well as a field called *searchterms*. The results of our runs are shown in Table I, with average precision figures when 5, 10, 15, and 20 documents were retrieved. BBNSLA and BBNSLB are the runs judged; the other runs are not.

Note that all our cross-lingual runs outperformed our monolingual run. This is not very surprising, given that similar results have been achieved in TREC [Xu and Weischedel 2001]. We are disappointed that query expansion did not significantly improve the results, unlike our prior experience with query expansion; in fact, query expansion sometimes hurts the results. However, in relevance assessments performed by

the National Institute of Standards and Technology, only runs BBNSLA and BBNSLB were used; therefore, the scores for runs BBNSLC, BBNSLD, and BBNSLMON are probably underestimates. Also, there are only 15 queries in the test set, making strong conclusions impossible.

6. RELATED WORK

The use of statistical language modeling techniques for IR and CLIR has appeared in a number of studies [Ponte and Croft 1998; Berger and Lafferty 1999; Miller et al. 1999; Hiemstra and de Jong 1999]. In particular, our CLIR model is similar to the one proposed by Hiemstra and de Jong [1999]. A difference is that our model makes use of the corpus statistics of the query language, while Hiemstra uses the corpus statistics of the document terms. The papers in this special issue are the first work we are aware of in porting CLIR techniques in such a limited time frame.

7. CONCLUSIONS

Overall, our cross-lingual retrieval system proved effective for Hindi in spite of the 29-day limit. All four cross-lingual runs achieved better results than our monolingual run. The multisite effort collected resources very quickly, making an effective CLIR system for a new language in 29 days possible..

REFERENCES

ALLAN, J., CALLAN, J., CROFT, W. B., BALLESTEROS, L., BYRD, D., SWAN, R., AND XU, J. 1998. INQUERY does battle with TREC-6. In *TREC6 Proceedings*. NIST Special Publications.

BERGER, A. AND LAFFERTY, J. 1999. Information retrieval as statistical translation. In *Proceedings of the ACM SIGIR Conference*. ACM, New York,1999.

HIEMSTRA, D. AND DE JONG, F. 1999. Disambiguation strategies for cross-language information retrieval. In *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries*. 274-293.

MILLER, D., LEEK, T., AND SCHWARTZ, R. 1999. A hidden Markov model information retrieval system. In *Proceedings of the ACM SIGIR Conference.* ACM, New York, 1999.

PONTE, J. AND CROFT, W.B. 1998. A language modeling approach to information retrieval. In *Proceedings of the* ACM *SIGIR Conference*. ACM, New York, 1998.

XU, J., WEISCHEDEL, R., AND NGUYEN, C. 2001. Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the ACM SIGIR Conference*. ACM, New York, 2001.

XU, J. AND WEISCHEDEL, R. 2001. TREC9 crosslingual retrieval at BBN. In *Proceedings of the TREC-9 Conference*. NIST Special Publications, 2001.

Received August 2003; revised September 2003; accepted December 2003